Theory Answers

1)
   a) The value of x can be within any range from (-inf,+inf) (as nothing about this is mentioned in the question). So the range of w_0+w_1x will also be between (-inf,+inf). As the range of the logistic function given domain=(-inf,+inf) is (0,1) our answer will be **(0,1)**
   b) The value of the logit function (log(x/(1-x))) is 0 when x=0.5, negative when x<0.5 and positive when x>0.5 also it is undefined (inf) when x=0 and x=1. From this, we can conclude that the range of logit function is (-inf,inf)

2)
   a) The MAE has an absolute term in it. Finding minimum optimizations for an absolute function is always harder than a regular function as they are not continuously differentiable functions and algebraic manipulations are required to compute them (for example if there is a correct prediction, MAE wont be differentiable at that point and can throw errors), which might translate to poor complexity while programming them. The MSE, on the other hand, is a simple square of terms function which is easily differentiable throughout and can be programmed easily without introducing extra complexity.

   b) The MAE is really helpful when there is a huge variance in the data. For example, let's say 10000,1,2,3,4,5. Clearly 10000 induces a high variance in the data and will increase the penalty for that particular example. But when we use MAE the penalty is not that huge (as compared to MSE). So when we have cases where we do not want to penalize the outlier examples a lot we use MAE. We also use MAE where we want a linear score for the errors (that is 10-0 is half of 20-0 and so on) which is not the case with RMSE and so on.
   c) Quantile loss is advantageous to MAE when there is a lot of variance in the data (either a lot of positive or negative variance). MAE is just

a particular case of Quantile regression (when quantile = 50%). Quantile loss is useful when we want to find the penalize underpredictions and over predictions by certain factors, rather than having a uniform penalty for both (which is what MAE does). In this way, we can optimize the potential prediction error. If we take lower quantiles (less than 50 %), then we penalize overestimates more (because there is a higher chance of the actual value being higher than the predicted value) and vice versa for higher percentiles. So we can take all the losses of different quantiles and find which quantile best describes the variance in the data and use that to make the model.

Programming Questions Report

1) i) c) The final results after convergence of a) are:

Average Train rmse= 2.533163606908052
Average Val rmse=2.5661062369136283

and for b) are:

Average training RMS Error=2.194948142213388
Average testing RMS Error=2.194948142213388

We can observe that the normal equations RMS errors are lesser than the normal linear regression.

ii)The optimization parameters keep changing (as I have randomized the creation of the folds).
Example parameters: Lasso hyperparameter {'alpha'= 0.0001}
Ridge hyperparameter {'alpha'= 0.1}

Test Rmse:
Test RMSE Error for l1 = 2.680226236741723
Test RMSE Error for l2 = 2.793198521532187

iii) The best fit line (of l1 and l2) visually moves away from the normal line, with l1 more closer to the normal than l2. This is mainly because the data has few features and overfitting is not possible in such a small dataset. And also as the best fit line is drawn on the train set itself, the regularization parameters reduce the accuracy on the train set by inducing some bias which makes the line look little inaccurate.

2) i) L1 is better than L2 regularization because it doesn't penalize outliers in a heavy manner as compared to L2 regularization. As L1 uses an absolute error and L2 uses a squared error the reasoning is similar to MAE vs MSE asked in the theory questions. Also, L1 is known to reduce coefficients to zero whereas L2 cannot do this. Given that we have 104 features in this question, This feature selection ability of L1 is particularly useful in eliminating some undesired features.

ii) The Train and test accuracy are:

Final Train accuracy for l1 regularization
0.9204166666666667
Final Test accuracy for l1 regularization
0.9183
Final Train accuracy for l2 regularization
0.9228666666666666
Final Test accuracy for l2 regularization
0.9177

As the train and Test accuracy are really close and the accuracy being very high, the model is a good fit which has low bias and low variance.